DEVELOPMENT OF ENVIRONMENTAL QUALITY PREDICTOR USING FEEDFORWARD ARTIFICIAL NEURAL NETWORK (FANN) IN MATLAB GRAPHICAL USER INTERFACE (GUI)

by

NAZIRA ANISA BINTI RAHIM

**Thesis submitted in fulfillment of the requirements
for the degree of
Master of Science**

May 2015

ACKNOWLEDGEMENT

Alhamdulillah. All praise to Allah S.W.T for His guidance and blessings to give me the opportunity, strength and inspiration to fulfill and complete my research thesis as planned.

My utmost gratitude to my supervisor, Associate Professor Dr Zainal Ahmad for his endless guides and advice during the studies. My most sincere thanks to all staffs at School of Chemical Engineering and Institute of Postgraduate Studies for their technical support and help during my studies here.

I would also like to extend my gratitude to the Universiti Sains Malaysia and The Ministry of Education for their financial supports through MyBrains. My gratitude also to The Department of Environmental (DOE) Malaysia and Mr Baharuddin Abdullah from The Department of Irrigation and Drainage Perak for their support in terms of research by sharing all the related primary data for modeling and analysis in Perak State .

Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ANFIS	Adaptive neuro fuzzy inference system
ANN	Artificial Neural Network
API	Air Pollutant Index
ASMA	Alam Sekitar Malaysia Sdn. Bhd.
BOD	Biological Oxygen Demand
BP	Back Propagation
CA	Correspondence Analysis
CAQM	Continuous Air Quality Monitoring
CCA	Canonical Correspondence Analysis
CCorA	Canonical Correlation Analysis
COD	Chemical Oxygen Demand
Coliform	Total coliform
Cond	Conductivity
DCA	Detrended Correspondence Analysis
DID	Department of Irrigation and Drainage
DO	Dissolved Oxygen
DOE	Department of Environment
DS	Dissolved solid
E-coli	Faecal coliform
FA	Factor Analysis
FANN	Feedforward Artificial Neural Networks
GRNN	Generalized Regression Neural Network

GUI	Graphical User Interface
HUM	Humidity
INWQS	Interim National Water Quality Standards
MAPE	Mean Absolute Percentage Error
MBAS	Methyl Blue Activated Substances
MI	Mutual Information
MIMO	Multi Input Multi Output
MISO	Multi Layer Single Output
MLP	Multiple Layer Perceptron
MPR	Multivariate Polynomiall Regression
MSE	Mean Square Error
OG	Oil and Grease
PCA	Principal Component Analysis
PCCA	Principal Canonical Correlation Analysis
pH	pH
PLS	Partial Least Square
PM_{10}	Particulate Matter
PSO	Particle Swarm Optimization
RBF	Radial Basis Function
RMSE	Root Mean Square Error
RSPM	Respirable Suspended Particulate Matter
RVS	Recursive Variable Selection
Sal	Salinity
SI_{DO}	Sub-index of DO
SI_{BOD}	Sub-index of BOD

SI_{COD}	Sub-index of COD
SI_{AN}	Sub-index AN
SI_{SS}	Sub-index of TSS
SI_{pH}	Sub-index of pH
SS	Suspended Solid
SSE	Sum Square Error
Temp	Temperature
TEMP	Temperature
T-Hard	Total Hardness
TS	Total Solid
Tur	Turbidity
WIND	Wind
WQI	Water Quality Index

PEMBANGUNAN PERAMAL KUALITI ALAM SEKITAR MENGGUNAKAN JARINGAN NEURAL SUAPAN HADAPAN DALAM ANTARA MUKA PENGGUNA GRAFIK MATLAB

ABSTRAK

Usaha pemuliharaan alam sekitar sentiasa berhadapan dengan kerumitan kerana ia melibatkan sejumlah besar pembolehubah. Walau bagaimanapun, dalam proses pembangunan model, pemilihan masukan yang betul untuk hasil ramalan yang berkaitan adalah penting. Tambahan pula, secara tradisinya laporan kualiti alam sekitar cenderung untuk lebih teknikal, menyampaikan data pemantauan alam sekitar yang tidak lengkap dan tidak mudah difahami. Oleh kerana data alam sekitar adalah berlebihan, kaedah pemilihan data input telah diperkenalkan; Analisis Koresponden Berkanun (CCA) dan Analisis Korelasi Berkanun (CCorA). Pendekatan-pendekatan ini boleh digunakan sebagai alat untuk memilih ciri dan bergabung dengan jaringan neural tiruan suapan hadapan (FANN) untuk membangunkan antara muka bergrafik (GUI) bagi peramal untuk para pengguna. Cadangan antara muka bergrafik untuk ramalan alam sekitar akan memberikan petunjuk tahap pencemaran air dan udara dan kualitinya, dengan terma-terma yang biasa digunakan oleh masyarakat. Untuk mencapai objektif tersebut, kajian ini telah dibahagikan kepada tiga fasa utama; penentuan pemilihan ciri masukan, pembangunan model FANN, dan akhir sekali, pembangunan GUI untuk pemantauan di luar talian. Terdapat dua kajian kes yang digunakan dalam kajian ini berdasarkan kepada data kualiti air sungai dan udara. Pengaplikasian CCA dan CCorA untuk menentukan masukan untuk ramalan telah berjaya dengan 7 (SS, NO_3, K, NH_3-NL, TS, Zn and Tur) dan 3 (kelembapan, suhu dan kelajuan angin) masukan pembolehubah telah dipilih untuk kajian kes 1 dan 2.

Keputusan menunjukkan bahawa rangkaian ramalan yang dibangunkan untuk sistem ramalan kualiti alam sekitar telah dilaksanakan dengan baik bagi masukan data yang sedikit. Secara umumnya, system ramalan yang dibangunkan berdasarkan FANN dengan kombinasi CCA dan CCorA telah menunjukkan prestasi yang baik dan membantu dalam memudahkan system ramalan alam sekitar ini. Model berbilang masukan – keluaran input-output tunggal telah berjaya digunakan untuk meramal indeks kualiti air (WQI) dan indeks pencemaran udara (API) dan berjaya dibangunkan dengan nilai regresi 0.90 dan 0.91 bagi kedua-dua rangkaian untuk data masukan yang belum pernah digunakan.

DEVELOPMENT OF ENVIRONMENTAL QUALITY PREDICTOR USING FEEDFORWARD ARTIFICIAL NEURAL NETWORK (FANN) IN MATLAB GRAPHICAL USER INTERFACE (GUI)

ABSTRACT

The environmental conservation efforts always deal with the complexity problem as it involves a large number of variables. However, in the process development of the model, the correct input selection for the corresponding output prediction is so important. Furthermore, traditional reports on the environmental quality tend to be too technical, presenting monitoring data without providing a complete and easy to understand facts of the environmental quality. Due to the redundancy of the environmental datas, input data selection methods were introduced; Canonical Correspondence Analysis (CCA) and Canonical Correlation Analysis (CCorA). These approaches could be applied as a feature selection tools and combined with Feedforward Artificial Neural Networks (FANN) to develop the graphical prediction interface for the end users. The proposed graphical user-interface for environmental prediction, will give an indication of the water and air pollution degree and their qualities, with the terms that are familiar within the community. To achieve those objectives, this research was divided into three main phases; determination of input feature selection, FANN model development and finally, GUI development for offline monitoring. Two case studies were used in this research which was based on river water and air quality data. The application of CCA and CCorA to determine the input for the prediction was successfully applied with 7 (SS, NO_3, K, NH3-NL, TS, Zn and Tur) and 3 (humidity, temperature and wind speed) input variables were selected for Case Study 1 and 2, respectively. The

results show that the developed prediction networks for the environmental quality prediction system has been executed well for less of input data. The developed prediction system based on FANN with the combination of CCA and CCorA generally has generally performed well and helped in simplifying the environmental prediction system. The final multi-input single output (MISO) models that have been used to predict the water quality index (WQI) and air pollution index (API) were successfully developed with the regression values of 0.90 and 0.91 for both of the networks for the unseen validation data input.

CHAPTER 1

INTRODUCTION

The environmental preservation efforts, especially on water quality and air quality had become a subject of ongoing concern. In the last decades, many researchers have monitored the gradual accumulation of reliable long-term quality data (Antonopoulos *et al.*, 2001). The environmental quality prediction had received more attention as it plays an important role in the control, management and planning of agriculture and aquaculture (Dhall *et al.*, 2008). Water quality is one of the main aspects in the environmental management and it will become the major constraining resource for sustainable development of large areas in the world.

According to Boyacioglu (2006), a basic problem in the case of water quality monitoring is the complexity associated with analyzing the large number of variables. Different multivariate statistical techniques such as principal component analysis (PCA) and factor analysis (FA) are used for the interpretation of a complex data (Iscen *et al*., 2008). Artificial Intelligence (AI) approach as predicting tools had also been applied for water and environmental quality studies (Li and Hassan, 2006).

One of the most popular AI methods that is currently being used is an Artificial Neural Network (ANN). These kind of models are inspired from the neurological system of humans and used to mimic the human neurological system. After all, it shows a remarkable success in the modeling and prediction of higher nonlinear systems including water quality prediction cases (Khuan *et al.*, 2002).

The greatest strength of neural network is it has the ability to learn the system from the history data. It has emerged out to be more flexible, less assumption dependent and adaptive methodology in environmental related areas such as water quality and air quality management (Boznar *et al*., 1993), lake and reservoir modeling, hydrologic forecasting and others.

Furthermore, a research by Rabiatul and Zainal (2012) supports that the main advantage of neural network based process models is that they are easy to build. This feature is particularly useful when modeling complicated processes where detailed mechanistic models are difficult to develop. However, the disadvantage of this model with a large set of functions is the degree of variables difficulties being estimated is increasing with the increasing amount of variables.

In the process development of the model, input variables must be chosen - based on the proven relationship to the outputs. According to Tarassenko (1998), the condition of input data somehow may affect the network; hence, reducing the input dimensionality is a must by limiting the existance number of free variables in the network. The free variables are the total number of biases and weights within the network and determined mainly by the number of inputs.

Therefore, the correct input selection for the corresponding output prediction is so important in neural network modelling. As much as it improves the prediction performance by reducing the number of data after eliminating input that less significant to the output, it is also provide faster and more cost-effective predictors (Guyon and Elisseeff, 2003).

The application of different multivariate statistical analysis techniques, like CCA and CCorA, for the understanding of the complex data matrices offers a better tool for reliable analysis of water and environmental quality system in respect to the trends and relationships of the data (Simeonov *et al.*, 2003; Kovac-Andric *et al.*, 2009). Canonical correlation analysis (CCorA) is a well-known technique to find maximally correlated projections between two data sets, which has been widely used in economics, meteorology and in many processing fields while canonical correspondence analysis (CCA) technique has been used to investigate distribution pattern of the environmental variables (Luca *et al.,* 2011). CCA also is extremely robust when assumptions of uni-modal responses do not hold with it also choses the best weight for the environmental variables and shows non-linear relations between species and environmental factors (Bodaghabadi et al., 2011).

Overall, the use of multivariate statistical analysis, such as CCA and CCorA in developing the environmental quality prediction interface has a potential to improve the prediction networks; therefore these approaches should be applied as a feature selection tool (Sakar and Kursun, 2012) and combine with FANN to develop the graphical prediction interface for the end users.

1.1 Problem Statement

In recent years, the development and current progress of utilizing the artificial intelligence into river waters and air quality modeling have been reported in many articles like Dawson and Wilby (2001), May *et al*. (2008) and Solomatine and Ostfeld (2008) and many more. Creating environmental quality prediction system with appropriate efficiency is challenging, especially in the evaluation of water and air quality. The problem of obtaining models that adequately represent the dynamic behavior of environmental data is also not easy. Khan *et al*. (2001) further added that lack of good understanding regarding the environment, the availability of reliable and complete field data set and the estimation of the numerous variables involves are the major factors contributing to this problem.

Since no individual variable can express the water quality sufficiently, the water quality is normally assessed by measuring a broad range of variables. However, increase in model complexity will undoubtedly increase the number of variables, leading to the problems of identification as stated in Beck (1986). This number, up to 30 variables should be reduced for the prediction system to function well and also to save the energy, time and manpower to collect all the variables. The reduced dimensions of the variables can also prevent the monitoring and prediction system from operating with an abundance of data. Since water quality is usually characterized by sets of physical, chemical and biological variables, which are mutually interrelated, only the highly contributed variables were needed for the system to operate, thus reducing the time taken by system to make the prediction. Thus, it is possible to significantly reduce the amount of input variance necessary for the neural network model, without considerably changing the predictive power of the model. This statement also apply to air quality prediction where it involves chemical and physical or meteorological data.

The increased concern about the environmental issues has encouraged researchers to focus on the proper monitoring, predicting and control of the environmental quality. Specific Graphical User Interface (GUI) for the prediction models for the end-user and the community could help them in understanding the meaning of the bunch of the data. Thus, a prediction model will become more effective if it is embedded into graphical user interfaces, especially for non-technical user. In fact, the program interface is taking an advantage of the computer's graphics capabilities to make the program user friendly. Well-designed graphical user interfaces can ease the users from learning complex command languages. Moreover, an easily understood and user friendly environmental prediction tool has insufficiently in the open literature so far (Chau, 2006).

1.2 The objectives of the thesis

The main objectives of the research work are defined as follows:

- To apply Canonical Correlation Analysis and Canonical Correspondence Analysis as a feature input selection approach.
- To develop the feedforward artificial neural network (FANN) for environmental prediction models for river water and air quality.
- To develop a hassle-free of Graphical User Interface that can be used directly to predict the water river and air quality status.

1.3 Scope of Study

This study was conducted to develop FANN for environmental prediction models for river water and air quality. It also looked into the application of multivariate statistical analysis techniques, CCA and CCorA for feature input selection. The study also developed an interface for final user to directly predict the water and air quality using the prediction system developed earlier.

The study was limited to an investigation using the river water quality data from Perak river basin and air quality data from 4 monitoring stations around Perak. The air quality data was collected for 5 years, from 2006 until 2010. All these data were collected by Department of Environment of Malaysia. The approach for using water quality index formulation has its limitations which are the disability in providing complete information on water quality and making full evaluation on all water risks. The formulation just provide a summary of the selected variables, thus the results can be subjective and biased, as it were affected by the formulation or based on the information gather from DOE.

1.4 Thesis Outline

This thesis is organized into five chapters; Chapter 1 to Chapter 5. In Chapter 1 presents an introduction regarding the research and the case study. A brief of problem statement and the research objectives are also presented in this chapter. Then, Chapter 2 presents a review of the related works of predicting and forecasting